Reb J H Close
University of California, Los Angeles (UCLA) Emergency Medicine Center
924 Westwood Blvd, Ste 300
Los Angeles, CA 90024
Carolyn J Sachs
UCLA Emergency Medicine Center
Los Angeles, CA
Pamela L Dyne
Department of Emergency Medicine
Olive View-UCLA Medical Center
Sylmar, CA

Correspondence to:
Dr Close
rclose@pol.net

Competing interests:
None declared

West J Med
2001;175:240-244

Reliability of bimanual pelvic examinations performed in emergency departments

ABSTRACT ● **Objective** To test the reliability of bimanual pelvic examinations performed in emergency departments by emergency medicine physicians. ● **Design** Prospective observational study; 2 examiners each recorded various pelvic examination findings on 186 patients. ● **Setting** A private university hospital and a public county hospital staffed by attending emergency medicine physicians who share an emergency medicine residency program. ● **Subjects** Senior resident (3rd or 4th year) and attending emergency physicians. ● **Main outcome measures** Percentage of agreement and percentage of positive agreement for cervical motion tenderness, uterine tenderness, adnexal tenderness, adnexal mass, and uterine size (within 2 cm). ● **Results** The agreement ranged between 71% and 84%, but the percentage of positive agreement was much lower, ranging from 17% to 33%. Agreement for uterine size, within 2 cm, was 60%. ● **Conclusion** The findings of bimanual pelvic examinations performed by emergency physicians in an emergency department have poor interexaminer reliability.

The bimanual pelvic examination is considered an important element in the evaluation of pelvic and abdominal complaints in women, including emergency conditions such as ectopic pregnancy, pelvic inflammatory disease, and tubo-ovarian abscess. Many sources list pelvic examination findings that would be expected with these disorders[1-4] but do not address the reliability or validity of the bimanual examination.

For a diagnostic test to have clinical utility, its results must be not only valid (accurate compared with the cri-

terion standard) but also reliable (reproducible when repeated, when done both by the same person and by different examiners).[5-10] A MEDLINE literature search from 1975 to the present failed to find any study of the reliability of bimanual examination among any physician group.

Although emergency physicians should not be expected to be as proficient at the pelvic examination as gynecologists, they are called on with great regularity—perhaps to a greater degree than almost any other nongynecologist provider—to make clinical decisions based on information derived from the pelvic examination. Thus, any deficiencies in information derived from emergency physician pelvic examinations are likely to have implications for other physicians, including primary care physicians. We designed this study to test the reliability of bimanual examinations performed by emergency physicians in emergency departments (EDs). A priori, a percentage of agreement on abnormal ("positive") findings of less than 50% defined poor reliability.

Summary points

- Bimanual pelvic examinations are commonly performed in the emergency department evaluation of women presenting with pelvic and abdominal symptoms
- The interexaminer reliability of this examination has not been evaluated
- We compared the findings of bimanual pelvic examinations performed in each consenting patient by 2 different physicians
- Bimanual pelvic examination findings in the emergency department showed poor interexaminer reliability

METHODS

This prospective observational study was performed on a convenience sample of 186 female patients who presented to 1 of 2 EDs because of pelvic or abdominal symptoms or both and who had a bimanual pelvic examination as part of the routine ED evaluation. The study was performed from August 1996 through August 1997 at a private university hospital and a public county hospital that are staffed by attending emergency medicine physicians who share a second- through fourth-year emergency medicine residency program. The hospitals have a combined annual ED census of more than 90,000 patients, and residents, who do a 2-week obstetrics rotation during their first residency year, perform about 200 bimanual examinations in the ED by the beginning of the third year of residency training, as 1 of us (P L D) has estimated.

Research assistants approached senior (3rd- and 4th-year) residents and attending emergency physicians who were preparing to do a pelvic examination as part of the routine evaluation of an ED patient and asked them to participate in the study. If the physician agreed, the patient was asked to participate. Eligible patients included any woman with a presenting complaint thought by the clinician to require a pelvic examination in the normal course of evaluation. Exclusion criteria included age younger than 18 years or having had a hysterectomy.

After the treating physician filled out a standardized data form regarding the bimanual examination, a second emergency physician with similar experience (3rd- or 4th-year resident or attending physician) was recruited to perform a second pelvic examination and to then complete a standardized data form. The 2 physicians were counseled not to communicate with each other regarding the examination before completing the data forms.

The patient's chief complaint and history of present illness were available to both examiners, and each physician could ask the patient any questions he or she thought would aid in the evaluation. No specific instruction was given to examiners about how to perform a bimanual examination, and definitions of abnormal (positive) and normal ("negative") findings were not provided. Rectovaginal examination was neither suggested nor recorded. Examiners could perform a speculum examination as desired, although this was not a measured aspect of the study. The time at which each physician examination occurred was recorded.

The data form included the evaluation of cervical motion tenderness, uterine tenderness, right adnexal tenderness, right adnexal mass, left adnexal tenderness, and left adnexal mass. For the purposes of this study, responses for each variable were limited to "clinically significant," "not clinically significant," and "unsure" because of examination limitations. Research assistants were instructed to clarify, if asked by the clinician, that "clinically significant" implied a finding about which the physician would then take some (diagnostic or therapeutic) action. Subjects were also asked to evaluate uterine size. For this variable, the physician could give a number in centimeters or answer "unsure" because of examination limitations. When a physician answered "unsure," he or she was asked to note the factors that limited the examination.

The percentages of agreement and of positive agreement were calculated for each variable. The percentage of agreement is defined as the number of observations in which the examiners agree, divided by the total number of observations. The percentage of positive agreement evaluates the subset of patients for whom 1 or both examiners reported an abnormality. It is defined as the number of observations in which the examiners agreed on an abnormal finding, divided by the total number of observations in which 1 or both examiners reported an abnormality.

Table 1 *Comparison of physician responses for each measured variable**

Physician responses	Physician 1 Yes	No	Unsure
Cervical motion tenderness			
Physician 2			
Yes	5	15	0
No	15	145	4
Unsure	2	0	0
Uterine tenderness			
Physician 2			
Yes	20	24	4
No	23	108	3
Unsure	2	2	0
Adnexal tenderness (left or right)			
Physician 2			
Yes	32	46	11
No	40	214	11
Unsure	4	11	3
Adnexal mass (left or right)			
Physician 2			
Yes	3	8	3
No	8	295	21
Unsure	2	22	10

*The numerals represent the number of patients for whom physicians reported each response.

For adnexal mass and adnexal tenderness, the results for the right and left sides were combined.

The study population was subdivided according to whether the examiners categorized the examination as limited. In addition, to evaluate whether the results of the examinations were found to be unreliable because of actual differences in the patient's examination over time, we divided subjects into 2 groups based on the interval between examinations. We grouped a response of "unsure" with responses of "clinically significant" because physicians are likely to continue the evaluation (ie, order an ultrasonogram, expedite follow-up) if uncertain about physical examination findings.

Uterine sizes determined by either examiner to be less than 12 cm were evaluated. Examiners' estimates that differed by more than 2 cm defined discordance.

The human subjects review board at both institutions approved the study with a waiver of written consent. Verbal informed consent was required for patient participation.

RESULTS

A total of 205 physicians were approached and agreed to participate in the study (100%). Seventeen patients declined enrollment, and 2 patients were not enrolled because a second examiner was unable to participate. The resulting convenience sample totaled 186 (91%) nonconsecutive, individual patients whose cases were evaluated by 2 examiners. Their mean age was 31 years, with a range of 18 to 85 years. Most (125 [67%]) were Hispanic, 36 (19%) were white, 15 (8%) were African American, 6 (3%) were Asian, and 4 (2%) were from other racial or ethnic groups.

Physicians' responses are compared in table 1, and table 2 lists the percentages of agreement and percentage of positive agreement for each variable measured. The percentage of agreement ranged from 71% to 84%. However, when at least 1 examiner recorded a clinically significant finding, the proportion of patients for whom the second examiner agreed with the finding was much lower, ranging between 17% and 33%.

One hundred thirty-six patients (73%) had a uterine size estimated by both examiners. In 127 of these patients (93%), 1 of the examiners estimated the size to be 12 cm or less. The differences in estimated size for individual patients ranged between 0 and 10 cm, and the second examiner's estimate was within 2 cm of the first examiner's estimate in only 76 patients (60%) (figure).

In 372 individual examinations, the physician identified the examination as limited in 137 (37%). About half the patients (95/186 [51%]) had at least 1 element of a limited examination, according to 1 or both physicians. "Unable to determine" was noted for each variable between 1.7% and 9.4% of responses. Obesity was identified as the limiting factor in 63% of these examinations, and pain, anxiety, and retroversion were identified as the limiting factor in 21%, 8%, and 6% of examinations, respectively.

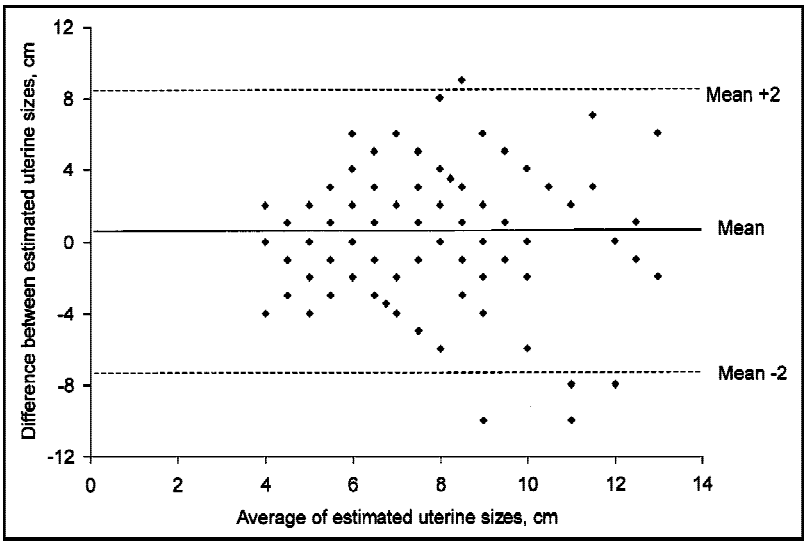
Table 2 *Percentage of agreement and percentage of positive agreement for measured variables*

Variable	Agreement, %	Positive agreement, %
Cervical motion tenderness	82	17
Uterine tenderness	72	33
Adnexal tenderness (right or left)	71	32
Adnexal mass (right or left)	84	23

We compared differences in the percentage of agreement and percentage of positive agreement between subjects for whom the examination was recorded as limited and those for whom it was not. As expected, the physicians appeared to be more reliable for those patients in whom neither thought the examination was limited, but percentage of positive agreement remained poor (<50% for all variables) and was not improved in 2 of the 4.

Research assistants recorded the time between physician examinations for 95% of study patients. Timing of the examinations did not have a large effect on reliability because the percentage of positive agreement for examinations performed within 20 minutes was statistically similar to those performed more than 20 minutes apart: 13% to 37% and 25% to 36%, respectively.

To evaluate the effect of multiple examinations on examination findings, we calculated the incidence of positive findings for each variable on first and second examinations. For each variable, there was no significant difference in the incidence of positive findings for first versus second examinations (table 3).



Comparison of estimated uterine sizes

DISCUSSION

In our study, interobserver reliability of bimanual pelvic examination performed by emergency physicians in EDs was poor. Emergency physicians perform such examinations and make patient care decisions based on examination findings regularly, perhaps to a degree greater than any other nongynecologic specialist. Results in this group of physicians should, therefore, be relevant to other generalist specialties. They should not be taken to mean that reliability is also poor among experts, such as gynecologists (although reliability of the bimanual examination among gynecologists has never been reported).

Many possible explanations can account for why bimanual examination performed unreliably among our subjects. First, examining physicians may have different definitions of what constitutes an adnexal mass or adnexal tenderness. Differences in examination findings may simply reflect inadequate medical school and residency teaching of pelvic examination. It may be possible to improve reliability by standardizing the instruction of bimanual examination and strictly defining abnormal and normal ("positive" and "negative") findings. For teaching purposes, examinations done by students and residents would need to be repeated or performed under direct supervision by attending physicians and the results discussed. Because more than 90% of the patients approached during this study consented to having a second examination, reexamination for teaching purposes is feasible. If the pelvic examination is indeed a good "test" to evaluate female pelvic symptoms, then reliability should be good after uniform teaching of the examination process.

Alternatively, the pelvic examination may be a poor test. This possible explanation is supported by literature that has found this examination to be inaccurate when performed by gynecologists.[11-14] Two studies have evaluated the validity of bimanual examination compared with ultrasonographic and laparotomy findings.[13,14] Disagreement between findings of bimanual examination and those of transvaginal sonography ranged from 9% to 54%. Only 29% agreement was found between findings of the bimanual examination and laparotomy. Even under ideal circumstances (pelvic examination under general anesthesia), the findings of bimanual examinations performed by attending gynecologists, gynecology residents, and medical students were inaccurate.[14]

Our study has several possible limitations. About half of our subjects were senior-level residents. Their experience may have been too limited, and a study using only attending physicians may yield different results. Pelvic ex-

Table 3 *Incidence of positive findings for first and second examination**

Variable	First examiner with positive finding, no. (%)	Second examiner with positive finding, no. (%)
Cervical motion tenderness	26 (14)	22 (12)
Uterine tenderness	52 (28)	52 (28)
Adnexal tenderness (right or left)	100 (27)	105 (28)
Adnexal mass (right or left)	46 (12)	48 (13)

*The denominator for cervical motion tenderness and uterine tenderness is 186, the number of patients. For adnexal tenderness and masses, the denominator is 372 because findings for the right and left sides were combined for analysis.

amination may be a skill that requires many years to learn, and the examination could be expected to be unreliable when performed by less experienced physicians. This is unlikely, given that prior research of pelvic examination validity under general anesthesia found poor accuracy for attending gynecologists, gynecology residents, and medical students. The differences among the 3 groups were not statistically significant.[14]

A second limitation involves our inclusion of "unsure" responses in the "clinically significant" group analysis. This was based on the assumption that a physician who is unsure about examination findings would have a lower threshold for performing additional studies and evaluation. This may not reflect what is actually done. Calculations were performed with "unsure" responses considered equivalent to "not clinically significant," and the percentage of positive agreement was still less than 50%.

A third limitation may be the availability of the patient's history to both examiners. As mentioned, both physicians had access to the history and could ask the patient any questions they thought were pertinent, but the physicians may have differed in their taking advantage of this access. Arguably, the treating physician took a more complete history and performed other components of the physical examination, thus biasing their examination findings.

CONCLUSION

The findings of bimanual pelvic examinations performed by emergency physicians in the ED are unreliable. Whether this is due to poor performance of the examination or because the examination itself is a poor test is unclear. The examination should not be abandoned, but physicians need to appreciate its limitations and not rely solely on the findings of a bimanual examination for clinical decisions.

References

1 Tintinalli JE, Ruiz E, Krome R. *Emergency Medicine: A Comprehensive Study Guide.* 4th ed. New York: McGraw-Hill; 1996.
2 Rosen P, Barkin R. *Emergency Medicine Concepts and Clinical Practice.* 4th ed. St Louis: Mosby-Yearbook; 1998.
3 Schwartz GR, Hanke BK, Mayer TA, et al. *Principles and Practice of Emergency Medicine.* 4th ed. Baltimore: Williams & Wilkins; 1998.
4 Dart RG, Kaplan B, Varaklis K. Predictive value of history and physical examination in patients with suspected ectopic pregnancy. *Ann Emerg Med* 1999;33:283-290.
5 Koran LM. The reliability of clinical methods, data and judgments (1st of 2 parts). *N Engl J Med* 1975;293:642-646.
6 Koran LM. The reliability of clinical methods, data and judgments (2nd of 2 parts). *N Engl J Med* 1975;293:695-701.
7 Gjorup T. Reliability of diagnostic tests. *Acta Obstet Gynecol Scand Suppl* 1997;166:9-14.
8 Fleiss JL. The measurement of interrater agreement. In: *Statistical Methods for Rates and Proportions.* 2nd ed. New York: John Wiley & Sons; 1981:212-236.
9 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304:1491-1494.
10 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
11 Andolf E, Jorgensen C. A prospective comparison of clinical ultrasound and operative examination of the female pelvis. *J Ultrasound Med* 1988;7:617-620.
12 Voss SC, Lacey CG, Pupkin M, Degefu S. Ultrasound and the pelvic mass. *J Reprod Med* 1983;28:833-837.
13 Carter J, Fowler J, Carson L, Carlson J, Twiggs LB. How accurate is the pelvic examination as compared to transvaginal sonography? a prospective, comparative study. *J Reprod Med* 1994;39:32-34.
14 Padilla L, Radosevich D, Milad M. Accuracy of the pelvic examination in detecting adnexal masses. *Obstet Gynecol* 2000;96,4:593-598.